Literature Review

Nicholas Chan

Index Terms—CSE185E, Literature Review, IEEEtran, Genomic Database, Nextstrain, UShER, Phylogeny.

INTRODUCTION

The database keeping and management of genomics data originated in the early 80's as a marriage of phylogeny and simple computer science algorithms for string parsing. In just the last year, all active genome databases have seen an unprecedented and catalytic growth due to the SARS-CoV2 pandemic. With an uptake of sequencing data equating to over half of the total number sequences submitted since 1982 occurring in just 2020[1], such growth is unlikely to falter until the pandemic is contained. Such rapid growth has been the result of a global effort focused on sequencing genomes of a single organism that, while impressive in scale, has been hastily assembled. My final paper on the main algorithms used for phylogenetic analyses and quality control will delve into the foundations of genomics database keeping and key pattern recognition in genomics. The backgrounds of these topics will be touched upon in this literature review.

1 GENOME DATABASES AND REAL TIME PATHOGEN TRACKING WITH NEXTSTRAIN

Nextstrain is considered as one of the progenitors of what modern viral genome databases were to become, a highly accessible database and web tool with a dynamic library of sequencing data. Similar databases and webtools have existed for a relatively long time prior to Nextstrain's arrival, but Nextstrain main claim to fame was this new concept of real time tracking of heavily sequenced pathogens. Nextstrain's proof of the viability of this concept lied in its humble beginnings as a visualizer for the spread and tracking the mutations of Influenza, West African Ebola, and Zika.[2] Soon after the dawn of the SARS-CoV2 pandemic in December 2019, Nextstrain became a centerpiece of both sequencing data queries and submissions.

Nextstrain Features

Queries run through Nextstrain take new genome sequences of an organism as input and output a phylogenetic tree built with a reference genome sequence as its root. With a figure as simple as this phylogenetic tree, a map of the spread of a virus can be drawn out on a global map, with the option to highlight variants in different colors. This feature has been coined as joint temporal and spatial visualization, because of Nextstrains ability to illustrate evolution through space(the world) and time. However, Nextstrain's role as a sequencing database has been reduced due to its partnership with GISAID, a more widely known public repository for sequencing data that has seen its own unprecedented uptake in SARS related datasets. Although many databases such as the UCSC Genome browser, NCBI 1000 Genomes Browser and Ensembl Genome Browser have existed for a considerable amount of time prior to Nextstrain's release, it was Nextstrain's openness to publicly sourced data in addition to professionally curated data that allowed it to skyrocket in a time where a dynamic database was in the utmost demand.

2 Genomic surveillance in the Beginning of the Pandemic

The report *Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California — Science*[3] by UCSF researchers investigated the epidemiology of SARS-CoV-2 and provided insight on the spread and evolution of the virus in Northern California from late January to mid-March 2020. In this article, Covid-19's spread to California was explained using the phylogenetic analyses tools and genome databases available on the Nextstrain site. Various Covid-19 genomes were collected to identify the lineage of the virus to track its intermediate variants through its dispersion from Wuhan China to California. Using methods such as Metagenomic Sequencing with Spike Primer Enrichment(MSSPE) and Tiled Multiplex PCR, genomes were recovered from NorCal COVID-19 patients, compared with the genomes on Nextstrain and then submitted to the GISAID database.

Results and Debunking of New Hypothesis

The comparative experiment on these Northern Californian variant sequences with the reference sequence of the Coronavirus (as well as the variant sequences known at the time of this study) provided the results necessary for precisely tracking the spread of the virus. Cases of SARS-CoV-2 across Northern California were found to be derived from 5 distinctive lineages. The most common was the Washington State(WA1) lineage which made its arrival to Northern California through a Grand Princess cruise ship. This Washington variant then spread to Sonoma, Marin, Solano, and San Joaquin County. The Santa Clara County Cluster and San Benito Cluster appeared from their respective counties. The European/New York lineage appeared in Sacramento, San Francisco, and San Mateo County. The San Benito and European/New York lineages were traced to international and interstate travel. The WA1 lineage of SARS-CoV-2 has been consistently linked to sequenced genomes of infected persons throughout the United States. These results showed that the virus from WA1, the first Case in Washington, was the main virus variant found in America. This can be visualized on Nextstrain by running the datasets used in this study. One of the significant factors in the increased spread of the NorCal variant was international travel to New York and the Grand Princess cruise, and intercontinental travels from places such as New York back to California. Following this study's conclusion on the true origin of the NorCal variant of Covid-19, a claim that SARS-CoV-2 virus like most other RNA viruses would have a slow rate of evolution was made simply as a hypothesis at the end of this study. Such a claim was viable at the time due to all the sequencing data available at the time. However only around half a year after this study by UCSF researchers, the rapidly mutating B.1.1.7 variant of SARS-CoV2 began to make its rounds in the UK. Such rapid transmission and evolution is free for anyone to see by running a query on the current reference phylogeny of SARS-CoV2 hosted on Nextstrain.

3 The Recent History of Virus Tracking Software

Professor Corbett-Detig's paper *Stability of SARS-CoV-2 phylogenies*[4] describes issue of preserving quality and accuracy of SARS CoV-2 sequencing data in this new fast-paced effort to keep up with the need for information on SARS-CoV2. Wetlab and bioinformatics associated errors have been identified as the main cause for reliability issues among sequencing data and have been strongly linked to recurrent mutations. To further confound results, the recurrence of these lab associated mutations causes them to appear less as outliers and more as legitimate mutations which coincidentially recur in nature. Although identifying and removing these problematic sites has been a challenge, the toolkit known as UShER has aided the effort in finding nodes where a mutations are likely to be erroneous and not caused as a result of evolution (UShER uses parsimony scores of nodes which we can make inferences from).

Features of UShER on the UCSC Genome Browser

UShER was developed by Professor Corbet-Detig's lab along with a Post Doctorate currently at UCSC. UShER is available for use through the UCSC Genome browser, where it has found a use for adding sequences onto previously composed phylogenetic trees in Newick format. The UCSC genome browser's port of UShER also has a pipeline directly to the Nextstrain website, where UShER's impressive speed as a phylogenetic tree builder is put to great use. UShER's implementation as a phylogenetic tree builder is based on the Fitch and Sankoff algorithms which have historically been used for phylogenetic tree construction with a maximum likelihood approach. Together, the Fitch-Sankoff algorithm work together in UShER to create acyclic trees where vertices (nodes) are mutations or points of divergence and edges are values of parsimony. The concept of maximum parsimony is interpreted in the context of computer algorithms and phylogeny as a metric for measuring minimal levels of character changes oberved in a tree. Trees with the lowest possible parsimony are the simplest and thus the most likely to not occur by chance.

REFERENCES

[1] National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1982] – [cited 2021 May 19]. [Online]. Available from: $https://www.ncbi.nlm.nih.gov/genbank/statistics$

[2] J. Hadfield et al. *Nextstrain: real-time tracking of pathogen evolution*, Oxford Academic, 2018.[Online]. Available: $https://pubmed.ncbi.nlm.nih.gov/29790939/$

[3] X. Deng et al. *Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California* San Francisco, California: AAAS, 2020. [Online]. Available: $https://science.sciencemag.org/content/369/6503/582$

[4] R. Corbett-Detig et al. *Stability of SARS-CoV-2 phylogenies*, PLOS Genetics, 2020. [Online]. Available: $https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009175$